Transformers

Transformers in natural language processing (NLP) are a type of deep learning module that uses self-attention mechanisms to analyze and process natural language data. They are encoder decoder models that can be used for many applications, including machine translation

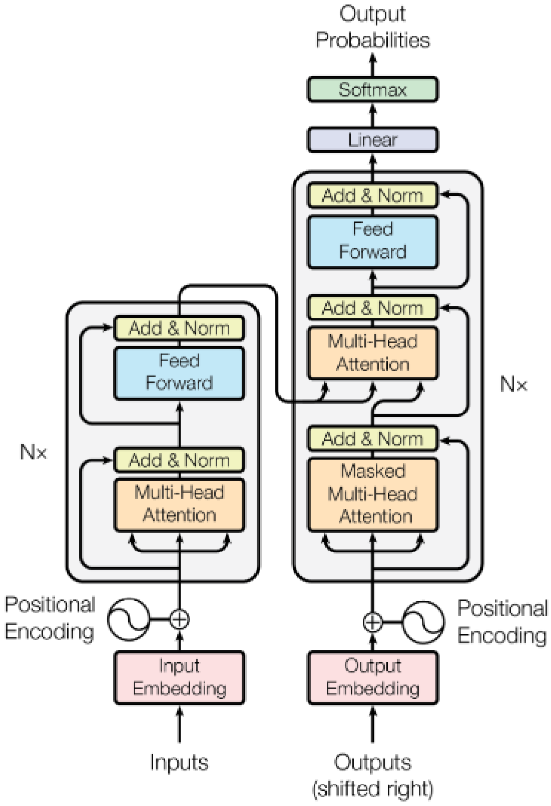
Attention Mechanism

Parallely we cannot send all the words in a sentence

Transformers

Using self attention to parallely send all the words

→ Architecture


Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)